[From: *Term banks for tomorrow's world: Translating and the Computer 4*.
Proceedings of a conference ... 11-12 November 1982, ed. Barbara Snell. (London: Aslib, 1983)]

TEXT PROCESSING IN DICTIONARY COMPILATION

Ken Moore
Longman Publishing Group, Harlow, United Kingdom

For some years now Longman have been both compiling and revising dictionaries with the aid of computers. Their use can easily be justified on word processing grounds alone, however there is scope for much more than this, for using the computer to aid the editorial processes, for revising the material in other publications and possibly for publishing the material in non-book forms.

A dictionary has a natural structure (headword, pronunciation guide, part of speech label, definition(s), run-ons etc) and if when the entry is keyboarded this structure is also captured then the material will be in a form suitable for text processing as well as for word processing. Clearly the cost of capturing this structure, together with a (relatively) small amount of additional coding (for example subject/semantic coding, a simple rating of the importance etc.), must not be neglected but the benefits can be considerable. The benefits fall into four categories reporting, checking, amending and reuse.

Reporting is used to establish the volume and balance of the material. Volume reports, numbers of entries, numbers of definitions, extent etc provide straightforward aids for management of schedules and resources. Additionally, though, these reports can be carried out on narrower areas of the material by making use of the coding, on subject or semantic field for example. This quickly identifies areas of weakness or excess and so helps the management of the balance of the material. Listing of problem areas, once identified can readily be produced and revised as individual activities.

Checking procedures are carried out in two forms, specific checking (defining vocabulary, or cross reference validity for example) and more general checking, identifying occurrences of a particular problem, scattered across the material as a whole perhaps.

Defining vocabulary is checked and controlled by usage counts of the vocabulary used and by listings of occurrences of specifically requested words in context. These two facilities enable both errors in vocabulary usage to be detected and corrected, and the use of infrequent vocabulary to be monitored and revised as necessary. This level of control is particularly useful in compiling language teaching courses and publications.

Cross Reference control can also be considerably simplified. Software is used to check firstly that each cross reference is satisfied, and secondly to make the cross referred material immediately available to the reviewer. This ensures substantially greater accuracy and helps eliminate circularity. It is particularly useful when the material is being cut back or revised extensively as is the case with derivative publications.

Lastly, useful work on derivative publications, simple checks on entry structure and sequence help prevent omissions and missequencing both within an entry and within a publication as a whole.

Checking procedures of the less specific type are based essentially on string searching techniques within the entry structure. Once a particular error has been recognised, and assuming it can be described in reasonably unambiguous terms, a listing of all occurrences of the error in context can readily be produced. In practice it can sometimes be difficult if not impossible to describe an error in exact terms and in such cases we would aim to produce excessive reports leaving the editor to make the finer judgement. Even in these circumstances, however, the work is likely to be very much less time consuming and more reliable than had it simply been done manually.

Amendment of text can be a difficult operation to carry out reliably by Software, however under careful control it can be a very powerful facility. It can be carried out at a simple level by ordinary string search and replace routines but for more complex amendments it may be necessary to develop specific software (although this can be costly).

Reutilisation of stored text is becoming increasingly important to us. The most obvious approach is dissimilar presentation of the same text for separate editions of a publication. Once the source material is in machine readable form the alternative product begins at the typesetting stage and requires no editorial and virtually no proof reading resources.

The production of derivative publications is a more interesting proposition with much more potential. Once the details of a new book have been established, and if a useful amount of material is already held on the computer system, a description of the material to be reused will be entered as parameters to a general extract program. A sample listing will be produced and checked. If necessary the extraction parameters will be amended, a new listing produced and so on until the most useful result is achieved. Using the final parameters, the new book will be created as a separate entity, completely listed and passed for editorial review. Here it may either be only cursorily checked, cut back, or expanded considerably, depending on the brief for the publication. The amendments will be marked directly on the listing and passed for keyboarding. The revised material will then be relisted, checked by an editor and so on until ready for publication. As a part of this overall process extensive use will be made of the reporting, checking and amendment facilities that I have previously described.

Finally the completed material is passed through a typesetting interface program which strips out all structure and coding, passing to the typesetter just the text and any running typographical commands necessary for automatic page make up.

Lastly, and perhaps by no means least, because the material is available in a computer readable form, because it is structured and coded for retrieval it is eminently suitable for publishing on other mediums than the paper for which it was first prepared. This is an area which offers considerable opportunity and which we are just beginning to explore.